Numerical Data in Science and Technology

David R. Lide, Jr.
National Bureau of Standards
Gaithersburg, MD 20899

Numerical data represent a particular subset of scientific information which is of major importance to both basic science and technology. Such data may describe the properties of individual entities such as atoms and molecules, of chemical compounds, or of materials such as alloys and ceramics. In these examples, the data are susceptible to measurement in the laboratory, and a new measurement should ideally yield the identical result regardless of when or where the measurement is done. Other data may refer to transient phenomena such as ocean currents, the appearance of sunspots, or astronomical events where the measurement is time dependent and cannot be repeated. In the traditional biological sciences, measurements on the characteristics of living organisms tend to show considerable variability, so that the statistical distribution over a population is the most useful type of information. However, in modern biology data are obtained which approach the level of precision of physics and chemistry. For example, the nucleotide sequences in DNA or the amino acid sequences of a protein can be determined in a reproducible way.

It can be argued that numerical data represent the most enduring component of the archival scientific literature. The theories, speculations, and explanations that appear in the literature tend to have a relatively short lifetime. If they are judged to be correct, then they are absorbed into the

mainstream of the science and it is rarely necessary to go back to the original literature to retrieve them. If they are not accepted, then the ideas are forgotten. On the other hand the numerical data that are published in the literature frequently have a very long lifetime. In some cases this may run into centuries or millennia. For example, astronomical observations recorded in ancient times are often of great interest to current astronomers. Thus it is important to the health of science and technology that data are preserved in a fashion that they can be retrieved and used by future generations.

We all know that the amount of scientific research carried out in the world has grown in an explosive way during the last generation. The production of data from this research has grown even faster because of the introduction of automated measuring instruments into the laboratory and the observatory. These data represent a large investment in both human and financial capital and should clearly be preserved for future use. We have already passed the stage where preservation in the form of printed documents such as journals, reports, and handbooks is feasible. Fortunately, modern computer storage technology is capable of preserving the scientific and technical data that are generated today. However, there are challenging problems in designing storage systems so that the contents can be identified, accessed, and used in an effective way.

A comment on the distinction between information and knowledge, as it applies to the type of scientific data under discussion, may be useful at this point. Any measurement that yields a quantitative value expressible in numbers is a piece of information. However, true knowledge is produced only

when the individual pieces of data have been correlated or fit through the use of some organizing principle. For example the wavelengths of the spectral lines emitted by atomic hydrogen, which are shown in the third column of Fig. 1, were measured with high accuracy a century ago. This constituted very reliable information that could be confirmed in different laboratories, but it does not seem appropriate to refer to this single set of numbers as representing knowledge. In fact true knowledge of the hydrogen atom only came later through the work of Niels Bohr, who developed a simple model which explained the observed data and allowed prediction of new wavelengths which could then be subjected to experimental tests. This knowledge of hydrogen is conveyed by the additional information linked to each wavelength value in Fig. 1. Our knowledge of the hydrogen atom was subsequently extended to deeper layers by physicists such as Schrodinger, Dirac, and Schwinger. This illustrates the fact that the word knowledge should always be used in a relative sense.

In planning for the storage of data it is important to focus on the knowledge aspect rather than treating data as sets of isolated bits of information. If an organizing principle can be found which permits data to be analyzed and correlated, very large sets of data can often be reduced to a simple algorithm or to a much smaller set of numbers. This not only simplifies the computer storage problems, but more importantly, it permits more efficient access to the data.

As one example, the properties of water are obviously of great interest in many areas of science and technology. Over the years, thousands of investigations of the physical and chemical properties of water and steam have

been made and tens of thousands of individual data points have been reported in the literature. While it would be possible to store all of these data points in a computer system, it would not be very useful to do so. However, it has proved possible to correlate all of the observed data on the thermophysical properties of water and steam by means of an equation of state based upon sound theoretical principles. The parameters in the equation are determined by optimizing the fit to all the observed data points. The form of such an equation of state is shown in Fig. 2. It is not exactly a simple equation, but one that is readily handled by even a modest computer of today. It is only necessary to store the form of the equation and the 50 to 100 coefficients which appear in it, and to couple these data with computational algorithms which can generate any desired property of water. This entire package fits easily onto one diskette which can be handled by a personal computer. Hence, someone interested in a given property of water at some specified temperature and pressure can ask the question and get an instantaneous answer from the system. It seems fair to say that this system contains true knowledge of water.

This is a particularly favorable example of the use of an organizing principle to correlate and compress a large amount of data, as the science of thermodynamics is highly developed and the theory permits many properties to be computed from a single basic equation. In certain other areas of physics and chemistry our understanding is not yet good enough to develop such models, and in the biosciences and geosciences the available tools are much more limited. Thus the quest for organizing principles which will allow us to simplify storage and retrieval of vast quantities of data will in some cases

require major advances in our understanding of nature. However there are many areas of science where the underlying theory is already well established and all that is required is to apply this theory to existing data sets. This is often very hard work, and to many scientists it is not very exciting, but it can be done with existing tools.

Let us now turn to improvements in the dissemination of numerical data. The development of modern computers and telecommunication systems has made a major impact on methods of accessing all types of scientific data. There are two approaches to computerized data dissemination, namely on-line networks and magnetic or optical media which can be delivered to local computers. On-line networks are already widely used for dissemination of bibliographic and other referral information, and we are now seeing the beginning of dissemination of numerical scientific data in this manner. Such networks make it possible to retrieve data from remote computers using local telephone connections; they provide instant access to many different (often geographically dispersed) data sources. On the other hand, tapes or disks sent by post to individuals or libraries provide a convenient method for accessing data by personal computers. It is possible to imagine replacing many shelves of books by a few diskettes which can be used with a personal computer.

An example of personal computer access to data is provided by a mass spectral database recently released by the National Bureau of Standards. This database contains mass spectra of about 45,000 compounds, and it has been incorporated into a search and display system suitable for personal computers with internal hard disk storage. The user can obtain essentially instan-

taneous display of the mass spectrum of any specified compound either in tabular or graphical form. The same information in hard copy form would require close to 10,000 pages and would take a large amount of shelf space. Furthermore, the computerized version permits access to the data by a number of different search techniques, some of which would be extremely tedious or impossible in the printed version.

Thus, computerized dissemination of scientific data is becoming a reality, and it seems likely that most scientists will make use of these methods within a few years. This trend raises a number of issues which both the scientific and the information communities need to address. First of all, intelligent design of the data structure and the retrieval software is essential if maximum utility is to be realized from data dissemination systems. Very often a database will be designed without adequate thought as to the future uses to which it will be put, and deficiencies become apparent after a great deal of money and effort has been expended. There is at present considerable duplication of effort on development of certain types of databases and, what is more serious, there is very little standardization in such areas as designation of materials, properties, error limits, and other important features of the data. The economics of computerized dissemination is still poorly understood, but there is a general perception that existing systems cost far too much for the average scientist to afford them. Although few political or legal barriers to data transfer have been noted up to now, many people are concerned that such barriers will develop in the future, particularly barriers to the transfer of data across national borders. Unfortunately, there is a tendency for political authorities to be suspicious

of data or other types of technical information when expressed in computerized form, even though the same information in conventional printed form would arouse no suspicion. There is a real need for education in the proper use of databases and other technical information sources. This is an area that is badly neglected in the higher educational systems in most countries. Finally, developing countries present a problem because of the high cost of current computerized information services. Lack of access to good technical information is a barrier to development, and many people fear that the high cost of current computerized information systems will accentuate the gap between poor and rich nations.

One organization which is attempting to address some of these issues concerning computerized storage and dissemination of numerical data is CODATA, the Committee on Data for Science and Technology of the International Council of Scientific Unions. CODATA has representation from both countries and scientific unions and thus attempts to address problems on both an international and interdisciplinary basis. CODATA activities fall into several categories. It supports publications and workshops which address problems in database design and software for manipulating data. CODATA also attempts to coordinate the efforts of data centers in different countries which are collecting the same type of data and in several cases has developed standard interchange formats which facilitate the transfer of data from one data bank to another. Through its international conferences, CODATA has attempted to air some of the potential economic, political, and legal problems. It has also developed various international directories to sources of data in different disciplines, and, in conjunction with Unesco, CODATA holds training

courses for technical information specialists in developing countries. By these means CODATA has provided a focus for many organizations and individuals concerned with numerical scientific data.

In summary, revolutionary changes are taking place in the way numerical data are stored, manipulated, and disseminated. Modern computer technology permits large amounts of data to be stored in an inexpensive manner. However, it would be a mistake simply to dump massive amounts of raw data into storage files. The more intelligent course is to evaluate, organize, and compress the data, using all available theoretical tools of the science in question. In this way true knowledge bases can be built up which will provide resources for both present and future generations of scientists.

Much remains to be learned about the most effective way to disseminate scientific data with the aid of modern computers and telecommunication systems. In principle, computers can provide the means to distribute data more cheaply and more effectively than can be done with traditional printed works. However, the present costs of obtaining data through computerized systems is extremely high. We need to use some imagination and ingenuity to make the most effective use of these tools of modern technology.

HYDROGEN I (H^{0+}), Z = 1
Ground State $1s(^2S_{1/2})$ (1 electron)
Ionization Potential 109 678.7737 cm^{-1}; 13.599 eV

Multiplet	Rel. Int.	λ_{vac} (in Å)	Levels (in 10^3 cm^{-1})	Configurations	Terms	J - J	Notes	References
17	2	914.576	0.0 – 109.3402627	1s – 18p	$g^2S - {}^2P^\circ$	$\frac{1}{2} - \frac{3}{2}$		309
16	2	914.919	0.0 – 109.2992665	1s – 17p	$g^2S - {}^2P^\circ$	$\frac{1}{2} - \frac{3}{2}$		309,488
15	2	915.329	0.0 – 109.2503457	1s – 16p	$g^2S - {}^2P^\circ$	$\frac{1}{2} - \frac{3}{2}$		309,488
14	3	915.824	0.0 – 109.1913178	1s – 15p	$g^2S - {}^2P^\circ$	$\frac{1}{2} - \frac{3}{2}$		309,488
13	3	916.429	0.0 – 109.1191942	1s – 14p	$g^2S - {}^2P^\circ$	$\frac{1}{2} - \frac{3}{2}$		309,488
12	4	917.181	0.0 – 109.0297939	1s – 13p	$g^2S - {}^2P^\circ$	$\frac{1}{2} - \frac{3}{2}$		309,488
11	5	918.129	0.0 – 108.9171238	1s – 12p	$g^2S - {}^2P^\circ$	$\frac{1}{2} - \frac{3}{2}$		309,488
10	7	919.351	0.0 – 108.7723455	1s – 11p	$g^2S - {}^2P^\circ$	$\frac{1}{2} - \frac{3}{2}$		309,488
9	9	920.963	0.0 – 108.5819954	1s – 10p	$g^2S - {}^2P^\circ$	$\frac{1}{2} - \frac{3}{2}$		309,488
8	10	923.150	0.0 – 108.3247262	1s – 9p	$g^2S - {}^2P^\circ$	$\frac{1}{2} - \frac{3}{2}$		309,488
7	20	926.226	0.0 – 107.9650569	1s – 8p	$g^2S - {}^2P^\circ$	$\frac{1}{2} - \frac{3}{2}$		309,488
6	30	930.748	0.0 – 107.4404490	1s – 7p	$g^2S - {}^2P^\circ$	$\frac{1}{2} - \frac{3}{2}$		309,488
5	40	937.803	0.0 – 106.6321640	1s – 6p	$g^2S - {}^2P^\circ$	$\frac{1}{2} - \frac{3}{2}$		309,488
4	70	949.743	0.0 – 105.2916540	1s – 5p	$g^2S - {}^2P^\circ$	$\frac{1}{2} - \frac{3}{2}$		309,488
3	130	972.537	0.0 – 102.8238962	1s – 4p	$g^2S - {}^2P^\circ$	$\frac{1}{2} - \frac{3}{2}$		309,488
2	300	1025.722	0.0 – 97.4923214	1s – 3p	$g^2S - {}^2P^\circ$	$\frac{1}{2} - \frac{3}{2}$		309,488
1	670	1215.6683	0.0 – 82.2592865	1s – 2p	$g^2S - {}^2P^\circ$	$\frac{1}{2} - \frac{3}{2}$	P	309,488
1	330	1215.6737	0.0 – 82.2589206	1s – 2p	$g^2S - {}^2P^\circ$	$\frac{1}{2} - \frac{1}{2}$	P	309,488

Fig. 1. Spectral Lines of the Hydrogen Atom

$$A(\rho, T) = A_{\text{base}}(\rho, T) + A_{\text{residual}}(\rho, T) + A_{\text{ideal gas}}(T)$$

$$A_{\text{base}}(\rho, T) = RT\left[-\ln(1-y) - \frac{\beta-1}{1-y} + \frac{\alpha+\beta+1}{2(1-y)^2} + 4y\left(\frac{\bar{B}}{b} - \gamma\right) - \frac{\alpha-\beta+3}{2} + \ln\frac{\rho RT}{P_0}\right]$$

$$A_{\text{residual}}(\rho, T) = \sum_{i=1}^{36} \frac{g_i}{k(i)}\left(\frac{T_0}{T}\right)^{l(i)}(1-e^{-\rho})^{k(i)} + \sum_{i=37}^{40} g_i \delta_i^{l(i)} \exp\left(-\alpha_i \delta_i^{k(i)} - \beta_i \tau_i^2\right)$$

$$A_{\text{ideal gas}}(T) = -RT\left[1 + \left(\frac{C_1}{T_R} + C_2\right)\ln T_R + \sum_{i=3}^{18} C_i T_R^{i-6}\right]$$

Table A.2. Coefficients for Residual Function

i	$k(i)$	$l(i)$	$g(i)$ (J g^{-1})	i	$k(i)$	$l(i)$	$g(i)$ (J g^{-1})
1	1	1	−530.62968529023	19	5	4	−1380257.7177877
2	1	2	2274.4901424408	20	5	6	−251099.14369001
3	1	4	787.79333020687	21	6	1	4656182.6115608
4	1	6	−69.830527374994	22	6	2	−7275277.3275387
5	2	1	17863.832875422	23	6	4	417742.46148294
6	2	2	−39514.731563338	24	6	6	1401635.8244614
7	2	4	33803.884280753	25	7	1	−3155523.1392127
8	2	6	−13855.050202703	26	7	2	4792966.6384584
9	3	1	−256374.36613260	27	7	4	409126.64781209
10	3	2	482125.75981415	28	7	6	−1362636.9388386
11	3	4	−341830.16969660	29	9	1	696252.20862664
12	3	6	122231.56417448	30	9	2	−1083490.0096447
13	4	1	1179743.3655832	31	9	4	−227228.27401688
14	4	2	−2173481.0110373	32	9	6	383654.86000660
15	4	4	1082995.2168620	33	3	0	6883.3257944332
16	4	6	−254419.98064049	34	3	3	21757.245522644
17	5	1	−3137777.4947767	35	1	3	−2662.7944829770
18	5	2	5291191.0757704	36	5	3	−70730.418082074

i	$k(i)$	$l(i)$	ρ_i (g cm^{-3})	T_i (K)	α_i	β_i	g_i (J g^{-1})
37	2	0	0.319	640.	34	20000	−0.225
38	2	2	0.319	640.	40	20000	−1.68
39	2	0	0.319	641.6	30	40000	0.055
40	4	0	1.55	270.	1050	25	−93.0

Fig. 2. Equation of State (Helmholtz Function) of Water and Some of the Coefficients